# CASE STUDY D: Fine-grained analysis of student data at California State University

John Whitmer's doctoral research project examined student activity on a course at California State University, Chico. This is a mid-sized campus in the North West of the state which had 14,640 full-time students in 2010. Whitmer, now Director for Platform Analytics and Research for Blackboard, analysed the complete population of 377 students taking the "Introduction to Comparative Religion" course in the 2010 Fall Semester. The course had recently been redesigned to integrate much deeper use of learning technology. Whitmer's research attempted a more fine-grained approach to the factors leading to student success than the cruder measures used by some other researchers.

## Key takeaway points

- Predictions of student success are more accurate when using multiple demographic variables than single ones
- Variables relating to a student's current effort (particularly use of the VLE) are much better predictors of success than their historical or demographic data
- Cleaning and transforming the data is a complex but essential process
- Predictive analytics is better performed on categories of usage (e.g. engagement or assessment) than at the level of individual tools
- *Total hits* is the strongest predictor of student success, with *Assessment hits* coming a close second

Whitmer believed that existing learning analytics initiatives tended to ignore most of the data in the VLE and there was minimal justification for why some variables were included and not others. He wanted to find answers to the following questions:

1. Is the frequency of use of the VLE related to academic achievement?
2. How relevant in this are the pedagogical functions of the VLE e.g. do tools which promote engagement such as forums have more impact than administrative tools such as the calendar?
3. Is VLE use a better predictor of success than demographic and other personal data?
4. Are students deemed to be academically at risk (due to e.g. ethnicity or socio-economic group) impacted by VLE use differently from those considered not to be at risk?

Whitmer points to other studies (such as Arnold, 2010 and Macfadyen & Dawson, 2010) where alerts were triggered if a student had a low level of VLE use, but he says that the pedagogical features deployed are largely ignored by these researchers. He groups VLE use measures into broader categories e.g. posting to

a discussion is an *engagement* activity. He then looks for examples of whether students who are more engaged in this way are more successful academically. Whitmer refers to the categorisation of VLE use proposed by Dawson & McWilliam (2008): administration (viewing of announcements and use of calendar), assessment (use of assessment tools and assignment submission), content (viewing of course materials) and engagement (with other students and instructors through discussion tools and email).

The study also examined nine student characteristic variables:

1. Gender
2. Whether the student's racial/ethnic group is under-represented in higher education
3. Income status (based on whether the student qualifies for a federal financial assistance)
4. High school grade point average
5. First in family to attend college
6. University college in which the student has their major
7. Enrolment status e.g. first year, continuing student, transfer from another university
8. Whether the student is both from an under-represented racial/ethnic group *and* has low income
9. Whether the student is both from an under-represented racial/ethnic group *and* is male

Other variables which measure student motivation or learning styles for example could increase the accuracy of the predictions. Whitmer did not analyse the *quality* of student interactions such as a "thoughtful" posting to a forum but says that the inclusion of such data could further refine the model.

Whitmer considers that this work is relevant for those responsible for implementing technology in universities because it analyses an assumption behind their work: that increased use of technology in learning enhances academic achievement. He also points out that earlier studies consistently found that combinations of student characteristics correlated much better with student success than single demographic variables do. In other words, if the student's ethnicity/race, income level and gender are used, the predictions should be considerably more accurate than simply using one of these variables. These researchers found that variables relating to the student's current studies such as their financial aid status and current grade point average are stronger in predicting success than their historical data.

| Variable | % Var. |
|---|---|
| HS GPA | 9% |
| URM and Pell-Eligibility Interaction | 7% |
| Under-Represented Minority | 4% |
| Enrollment Status | 3% |
| URM and Gender Interaction | 2% |
| Pell Eligible | 2% |
| First in Family to Attend College | 1% |
| Mean value all significant variables | *4%* |
| | |
| **Not Statistically Significant** | |
| Gender | |
| Major-College | |

**Table 1: Predictive value of demographic / educational variables (Source Whitmer presentation to UK LA Network)**

# Collecting and transforming the data

Data was gathered from the VLE and from the student information system, and transferred to a data warehouse. Whitmer details a number of processes to reduce and transform the data:

1. **Clean the VLE logs.** First of all VLE log file records were filtered to remove any staff use from the analysis. 2% of records were found to relate to use of the VLE by staff. It was also found that one discreet action by a student could result in twenty hits recorded in the logs so these were consolidated where appropriate. Logs of activities with a dwell time of less than five seconds were then removed as these might represent automated server-level events. Activities with a dwell time of an hour or more were excluded too as these may have resulted from a student starting an activity and then leaving their machine. Use of tools which were rarely used by students were also discounted. By this stage 74% of records had been removed from the log files, a finding which Whitmer says is significant for future studies which consider VLE data for predicting student success. Without such filtering, he says, there is the potential for serious inaccuracies in the analytics.
2. **Clean the student characteristics data.** The six students who did not receive a final grade or received a "withdraw" grade were removed from the analysis.
3. **Join the datasets.** An anonymised identifier was set for each student to join the VLE log files to the characteristics and grade data.
4. **Transform and reduce the data.** For the subsequent regression process, some of the variables needed to be recoded in numerical format e.g. male=0, female=1. The *generated variables* such as under-represented minority (URM: no=0, yes=1) were created at this point from other data fields. The *interaction variables* were also calculated e.g. URM and Male (no=0, yes=1). Next the *aggregated VLE use variables* were generated for the different use categories of administration, assessment, content

and engagement so that the number of logs for activities in each of these areas could be included. Finally the data was consolidated into one record per student.

5. **Check for missing data.** Several variables were excluded from the analysis because they were missing in more than 10% of cases. For High School GPA, 5.63% of the variables were missing, for example. For this variable the mean over all the cases was used to impute the missing values.

6. **Carry out final checks.** Some final checks necessary for the statistical analysis were then carried out. The values were inspected to ensure they were sufficiently dispersed and also that there were no serious "outliers". Tableau data visualisation software was used for this. Variables were also checked to make sure they were independent from each other, as dependent variables would skew the analysis. Any with a correlation of greater than 0.5 would have been removed from the analysis but none were found.

# Findings

Seven of the nine student characteristic variables were found to be statistically significant. The VLE variable correlations however were more significant. In other words VLE use is a better predictor of student success than the demographic and enrolment data traditionally used to identify at-risk students. It was also found that use of individual tools varied widely between students but that there was less dispersed use within the categories which had been defined i.e. administration, assessment, content and engagement. Whitmer recommends therefore that predictive analytics should not be performed at the level of individual tools.

| VLE Use Variables (ordered by r values) | r | % Variance |
|---|---|---|
| Total Hits | .48 | 23% |
| Assessment Activity Hits | .47 | 21% |
| VLE Content Activity Hits | .41 | 17% |
| Engagement Activity Hits | .40 | 16% |
| Administrative Activity Hits | .35 | 12% |
| *Mean value all significant variables* | | *18%* |

**Table 2: Correlation results of VLE use with course grade (Whitmer thesis p.90)**

From Table 2 it can be seen that *Total Hits* is the strongest predictor of student success, with *Assessment Activity Hits* coming a close second. In his conclusions Whitmer suggests that using total VLE hits would be an excellent starting point for predictive modelling and early warning systems.

Whitmer was surprised that *Engagement Activity Hits* came below *VLE Content Activity Hits* but suggested that this was because much of the content was only available through the VLE rather than through lectures so it was essential to use this feature. Administrative Activity Hits have the lowest correlation coefficient which he points out is not surprising as viewing your calendar is likely to have less impact on achievement than submitting an assessment.

A multivariate regression was carried out using all the above variables except Total Hits. It was found that these explained 25% of the variation in final grade. Including all the student characteristic variables in the analysis added another 10% to the predictive relationship with course grade.

Overall Whitmer made the not-unexpected finding that use of the VLE is related to student achievement. What his research does do though is to *quantify* the difference the various types of VLE usage can make, and therefore enable instructors to monitor the efforts of their students and have at-risk students flagged to them. He found that VLE variables were more than four times as strongly related to achievement as demographic ones. VLE use is a proxy for student effort which is why it is such a strong predictor of final grade. It suggests that what students do on a course is more important than their background or previous results.

Another finding was that VLE use appears to be less effective for at-risk students: there was a 25% reduction in impact on final grade from VLE use by at-risk students (based on under-represented minority status and income) compared with not at-risk students.

There was widespread acceptance of this work by stakeholders from across the institution including administrators, researchers, faculty and student affairs. There were however several barriers to be overcome. The project was carefully examined in the light of the Family Educational Rights and Privacy Act (FERPA) and the local security policies. This resulted in the use of the anonymous identifiers and avoidance of the use of personally identifiable data. Finding time in the stakeholders' diaries to engage with the project was also challenging.

# References

Agnihotri, L., Ott, A. (2014) Building a Student At-Risk Model: An End-to-End Perspective From User to Data Scientist. *Proceedings of the 7th International Conference on Educational Data Mining.*

Arnold, K. (2010, March 3) Signals: Applying Academic Analytics. *EDUCAUSE Review*. http://er.educause.edu/articles/2010/3/signals-applying-academic-analytics

Dawson, S., & McWilliam, E. (2008) Investigating the application of IT generated data as an indicator of learning and teaching performance. Canberra.

Macfadyen, L. P., & Dawson, S. (2010) Mining LMS data to develop an "early warning system" for educators: A proof of concept. Computers and Education, 54, 588-599.

Whitmer, J. (2015) Using Learning Analytics to Assess Innovation and Improve Student Achievement. UK Learning Analytics Network. Nottingham. http://analytics.jiscinvolve.org/wp/2015/07/01/notes-from-uk-learning-analytics-network-event-in-nottingham/

Whitmer, J. C. (2012) Logging On to Improve Achievement: Evaluating the Relationship between Use of the Learning Management System, Student Characteristics, and Academic Achievement in a Hybrid Large Enrollment Undergraduate Course. University of California, Davis. https://johnwhitmerdotnet.files.wordpress.com/2013/01/jwhitmer_dissertation_complete_1-21-2013.pdf

Whitmer, J., Fernandes, K., & Allen, W. R. (2012, Aug 13) Analytics in Progress: Technology Use, Student Characteristics, and Student Achievement. Educause Review.
http://er.educause.edu/articles/2012/8/analytics-in-progress-technology-use-student-characteristics-and-student-achievement